Examples of Applying the Bilingual Dictionary to the Translation Algorithm in System POLENG

Przykłady zastosowania słownika dwujęzycznego w algorytmie tłumaczenia systemu POLENG

Krzysztof Jassem

Uniwersytet im. Adama Mickiewicza
Wydział Matematyki i Informatyki
ul. Matejki 48/49
e-mail: jassem@main.amu.edu.pl

ABSTRACT

The structure of the dictionary in the system POLENG (described in [1]) was designed with a view to implementing in the translation algorithm all information included in the dictionary. This paper shows how specific fields of description in the dictionary are applied to the algorithm. The interaction between the algorithm and the dictionary is exemplified by lexemes of the highest frequency occurrence in a corpus of computer-oriented texts.

STRESZCZENIE

Strukturę słownika systemu POLENG (opisaną w [1]) opracowano w taki sposób, aby informacja zawarta w słowniku mogła być w całości wykorzystana w algorytmie translacji. W poniższej pracy przedstawiono, w jaki sposób poszczególne pola opisu w słowniku znajdują zastosowanie w algorytmie. Współzależności między algorytmem i słownikiem są zobrazowane przykładami leksemów o najwyższej częstości wystąpień w korpusie tekstów komputerowych.

1. Preface

An electronic dictionary is understood by Natural Language Processing (NLP) researchers to be a dictionary readable by a computer. This means that the format of the dictionary should enable its automatic lookup and that the description of the entries should be 'computer-parsable' (we say that an expression is 'computer-parsable' if a computer procedure is able to state its correctness according to some rules). It seems that while creating an electronic dictionary yet another feature of the lexicon should be kept in view viz its applicability.

The format of the entries into the bilingual dictionary presented in [1] (and revised here) aims at its applicability in MT systems. A computer program which uses the information stored in the dictionary in order to translate Polish texts into English has been developed. The assumption has been made that the program should be able to take advantage of all types of information included in the dictionary. Below, the interaction between the translation algorithm used in the

program and the format of the dictionary is pointed out. For each attribute of the description, the way in which the algorithm may take advantage of its value is discussed. The examples shown in the paper originate from a dictionary consisting of lexemes occurring most frequently in Polish computer-oriented texts.

For the sake of completeness the format of an entry into the dictionary is revised in *Fig.1*.

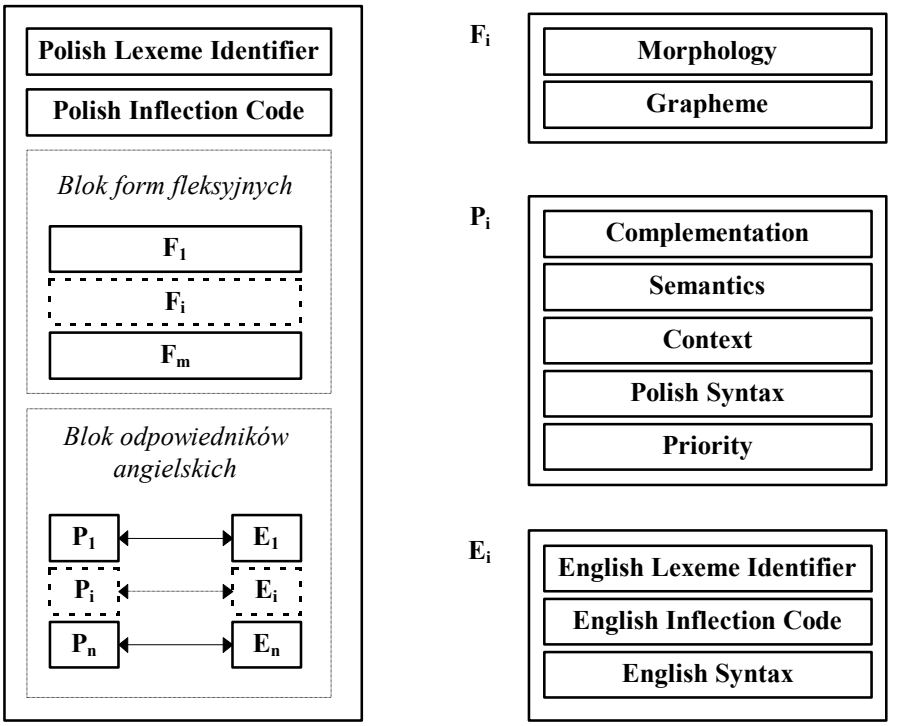

Fig. 1

2. Separating entries

An entry into the dictionary is identified by its canonical form (represented in *Fig. 1* as *Polish Lexeme Identifier*) and its inflection code (*Polish Inflection Code*). Note that the value of *Polish Inflection Code* fulfils two functions: first, it defines part of speech the lexeme represents and second, it enables automatic generation of inflected forms for flexional lexemes.

The condition for separating entries is defined as follows:

Two entries must differ either by the values of their *Polish Lexeme Identifier* field or by the values of their *Polish Inflection Code* field.

The first member of the above alternative is straightforward. Below, some examples of lexemes which are distinguished by the second element of the

alternative, are given. For each distinction its consequences to the translation algorithm are shown.

2.1. Separating entries which represent different parts of speech

Specifying that one canonical form may represent two different parts of speech is essential for the translation algorithm in:

- syntactical analysis of Polish text,
- determining the most adequate English equivalent.

It is necessary to distinguish between two functions of the word *wewnątrz* (English: *inside)*: as a preposition and as an adverb in order for the algorithm to be able to parse both following expressions: 1) *Sygnał przepływa wewnątrz systemu.* (*The signal is flowing inside the system*), *2) Sygnał przepływa wewnątrz.* (*The signal is flowing inside*).

The same takes place for other prepositions – adverbs, like: *powyżej (above), obok (near), niedaleko (not far away).*

The word *brak* behaves in some sentences as a regular Polish noun, e.g. *Ten sprzęt ma braki (This equipment has defects)* but it may also fulfil the function of a modal verb e.g. *Brak mi pieniędzy (I am short of money).* This distinction must be coded in the dictionary (by separating entries) so that the algorithm should be able to handle both functions of the word.

The word *choć* may function either as a conjunction or a particle (some sources do not distinguish between particles and adverbs, e.g. [2]; such an approach seems suitable for NLP needs). If an algorithm recognises *choć* as a conjunction, it should translate the word into *although*, whereas the English equivalent of the particle-adverb is *at least.*

There exist forms which represent more than two lexemes. The form *cokolwiek* represents the lexeme of a nominal pronoun (including inflected forms: *czegokolwiek, etc.*) having the English equivalent *anything*, the lexeme of a conjunction (English: *whatever*) and the lexeme of an adverbial pronoun (English: *somewhat*).

2.2. Separating entries which represent the same part of speech

Some entries are separated in the dictionary although they are represented by the same canonical form and belong to the same part of speech. This concerns mainly nouns and (rarely) verbs. The entries are distinguished by their inflection code.

The word *operator* (*English: operator*) belongs to different inflection paradigms depending whether it is understood as human or non-human. The differences occur in the nominative plural (*operatorzy* vs *operatory*) and in the accusative plural (*operatorów, operatory*). Creating separate entries for each meaning seems necessary: not only do their inflected forms differ but there exist differences in the morphological description of inflected forms (field *Morphology*) as well.

Several examples of this kind have been found in a corpus of computer-oriented texts: *przewodnik (human: guide, non-human: conductor), menedżer (menager), kontroler (controller).*

Two nouns are regarded as belonging to different inflection paradigms if subsets of their inflected forms are identical but one of them is defective (i.e. does not have either singular or plural forms). The lexeme *wzgląd* has several English equivalents. Some of them concern only defective lexemes. The equivalent *favour* may be viewed as corresponding only to the plural occurrences, the equivalent *consideration* would be the equivalent of only singular forms and the noun *respect* would correspond to both singular and plural forms. The dictionary contains three different entries for the noun *wzgląd* (*plural: względy*) including: only singular forms, only plural forms, all inflected forms.

Similar situation arises with the noun *życzenie*. Its English equivalents are *desire* (only singular forms), *greetings* (only plural forms), *wishes* (all forms).

This distinction does not ensure the determinism of the translation procedure. If a text contains a plural form of the lexeme *życzenia*, the algorithm cannot tell whether its equivalent is *wishes* or *greetings* (this distinction may optionally be made on the basis of other fields). The only benefit for the algorithm is that it may eliminate the equivalent *desire* as corresponding only to singular forms.

One infinitive may occasionally represent different verb lexemes. This is the case for the infinitive *stać (się) (English: to stand, to happen)* which belongs to two different lexemes, one of them being a perfect verb, the other being imperfect. Separating such entries is essential not only for generating proper inflected forms (most of them coincide for both lexemes) but also for assigning proper morphological information to the forms (consulted by the algorithm in the morpho-syntactical analysis).

3. Frames of inflected forms

Each entry includes a number of frames of inflected forms, marked in *Fig. 1* as F_i (the number of frames may vary from 1 for uninflected parts of speech up to over 50 for verbs). Participles and gerunds are treated as inflected forms of verbs, adverbs derived from adjectives are treated as inflected forms of adjectives (some of such derivatives are also inserted into the dictionary as separate entries if they do not obey compositional transfer). The field *Grapheme* stores the orthographic form of a word and is necessary in lexical analysis which is the first phase of the translation algorithm. The field *Morphology* includes information on grammatical features of an inflected form and is used in the morpho-syntactical analysis of the input text.

4. Application of the *Complementation* field

For a given entry the *Complementation* field describes grammatical categories and semantic features of the phrases which in a sentence are dependent on the entry. The discussion on what information should be included in the field may be

found in [1] and [3]. Here, we will have a look at how the translation algorithm may take advantage of the value of the field.

There are three phases in the translation algorithm in which the value *Complementation* is relevant:

- syntactical-semantic analysis,
- choosing the English equivalent,
- syntactical synthesis of English text.

These three phases will not be discussed here separately. Instead, the treatment of the field for various parts of speech will be described.

4.1. Verbs

The relevance of this field for verbs was described in [1] and [4]. As compared to the formalism presented in [4], new semantic features have been added. Below, a few examples of the treatment of this field for the verbs which occur most frequently in computer-oriented texts are shown in *Table 1*.

Polish Verb	Complement.	English Verb	Description (Examples)
zauważyć	TH → TH	notice	*zauważyć, że...* is transferred into *to notice that...*
	A → NP	spot	*zauważyć coś → to spot sth*
	OB → OB	observe	*zauważyć jak (kiedy...) → to observe (how, when)*
	DS → DS	remark	*"Pech", zauważył Jacek → "Bad luck", Jack remarked.*

Table 1

Polish Verb	Complement.	English Verb	Description (Examples)
skorzystać	z G: -Ab → NP	use	*skorzystać z czegoś (non-abstract complement) → to use sth*
	z G: Ab → of NP	take advantage	*skorzystać z ...(abstract complement) → to take advantage of sth*

Table 2

Polish Verb	Complement.	English Verb	Description (Examples)
odwoływać	A: Hum → NP	dismiss	*odwoływać kogoś (human complement) → to dismiss sb.*
	A: -Hum → NP	cancel	*odwoływać coś (non-human complement) → to cancel sth*

Table 3

Polish Verb	Complement.	English Verb	Description (Examples)
uważać	TH → RC	think	*"On, uważa, że to zrobię" → "He thinks I shall do it"* In this usage English verb is stative (does not occur in continuous forms), which is marked in *EnglishSyntax*
	na A → NP	mind	*uważać na coś → to mind sth*
	<A, za A> → <NP, as NP>	consider	*"On uważa mnie za głupka" → "He considers me as a fool"* (stative usage)
		be careful	*uważać (no complements) → to be careful*

Table 4

4.2. Nouns

The task of assigning complementation to nouns is even harder than that for verbs. The reason is that available standard dictionaries tend to show less interest in describing noun complements than verb complements.

Practise has shown that for nouns two types of complements need description: prepositional phrases and nominal groups in genitive.

An old-age problem in Natural Language Processing is determining whether in a given sentence a prepositional phrase (PP) modifiers a noun or a verb (as in the sentence: *Zatelefonował do mnie kolega z Francji,* English*: I was phoned by my friend from France*). The solution assumed in the POLENG system is that a PP modifies a verb unless stated as a noun modifier in the dictionary (thus, in the above sentence the system would assume the PP to modify the verb, as it is hard to imagine the dictionary to include the information that a *friend* may be modified by a PP: *z + Genitive*). A lexicographer should keep this rule in mind when deciding what kind of PPs to describe in the field. Beneath two examples from the dictionary are presented:

Polish Noun	Complement.	English Noun	Description (Examples)
odwołanie	od G -> from NP	appeal	
	do G → to NP	call	This translation proves correct in most cases in computer-oriented texts
	G:Hum → of NP	dismissal	*Odwołanie kogoś* (human complement)
	G:-Hum → of NP	cancellation	*Odwołanie czegoś* (non-human complement)

Table 5

Polish Noun	Complement.	English Noun	Description (Examples)
znaczenie	dla G → to NP	importance	*znaczenie dla kogoś → importance to sb.*
	no complement	meaning	

Table 6

The algorithm would look for complements in the neighbourhood of nouns they modify (failing if a modifier is separated from the noun: the failure would result in assigning the modifier to a verb).

4.3. Adjectives

Complements of adjectives (as treated here) are either prepositional phrases or nominal groups in a given case. The algorithm would look for complements in the neighbourhood of adjectives they modify. Two examples of adjectives and their complements are given below:

Polish Adj.	Complement.	English Adj.	Description (Examples)
właściwy	do G → to NP	fit	*właściwy do tego zadania → fit to the task*
	D → of NP	characteristic	*właściwy dobremu sprętowi → characteristic of good equipment*
	no complements	proper, suitable, appropriate, just, exact	these equivalents are distinguished by the *Priority field*

Table 7

Polish Adj.	Complement.	English Adj.	Description (Examples)
typowy	dla G → of NP	typical	*typowy dla tego niego → typical of him*
	no complements	standard	

Table 8

4.4. Prepositions

The information in the *Complementation field* for prepositions concerns nominal groups that may follow the prepositions. For some prepositions a lexicographer decides that it is an advantage to separate gerunds from other types of nominal groups, e.g. *dla G_GR → to IN (dla skonstruowania komputera → to design a computer).* Frequently the information allowed to be stored by the formalism does not suffice to determine the proper choice of the English equivalent. The equivalents are listed in the dictionary (distinguished only by the *Priority* field), and the system makes it possible for the user to choose a better equivalent from the list.

Note that in order to translate a PP the algorithm first checks for the PP to fit as a complement of other parts of the sentence and only when this fails, does the algorithm look for a preposition in the dictionary.

Polish Prep.	Complement.	English Prep.	Description (Examples)
przed	I: -Ab	in front of, ahead of, before	*przed budynkiem → in front of the building*
	I: Ab	before, against	*przed katastrofą → before (against) the catastrophe*
	A	in front of, ahead of, before	*przed budynek → in front of the building*

Table 9

Polish Prep.	Complement.	English Prep.	Description (Examples)
pod	I → NP	under, below, beneath, underneath, near, at	*pod drzewem → under the tree*
	A → NP	under, below, against	*pod drzewo → under (against) the tree*

Table 10

4.5. Adverbs

The *Complementation* field for adverbs is in most cases treated by the algorithm in the same way as for adjectives, e.g.

Polish Adv.	Complement.	English Adv.	Description (Examples)
niezależnie	od G → of NP	irrespective	*niezależnie od komputera → irrespective of the computer*
	no complements	independently	*zrobili to niezależnie → they did it independently*

Table 11

However, the value of the field may also distinguish between different functions of adverbs in a sentence:

Polish Adv.	Complement.	English Adv.	Description (Examples)
mało	(AJ → AJ | AV → AV)	not very	*mało przekonujący → not very convincing*
	N → NP	hardly	*mało kto → hardly anyone*
	G → NP	little, a few	*mało pieniędzy → little money*

Table 12

This solution is, however, controversial. An alternative approach looks attractive viz. to classify adverbs according to their syntactical features and to make a distinction between adverbs in the *Polish Inflection Code* field.

4.6. Adverbial Pronouns

Polish Ad Pr	Complement.	English Adj	Description (Examples)
tak	(AJ → AJ | AV → AV)	so	*tak (dobry, dobrze) → so good (well)*
	N → NP	such	*tak dobry człowiek → such a good man*
	no complements	this way	*zrób to tak → do it this way*

Table 13

This solution is also controversial (for the same reasons as for adverbs).

5. Application of the *Semantics* field

All nouns (and nominal pronouns) in the dictionary are assigned their semantic value. The classification consists of four exclusive sets: human, animate

(non-human), abstract, non-abstract (inanimate). The translation algorithm assigns the semantic value to a nominal group on the basis of the information stored in the dictionary under the heading of the basic noun. This mapping is essential for disambiguation of complements. Complements may be assigned two additional features: -Hum (non-human) and -Anim (inanimate, abstract or non-abstract). We could see how this specification is dealt with by the algorithm in the previous paragraphs. It is worth noting here that the *Semantics* field may be non-empty also for verbs. If the field is non-empty, then it denotes the semantic feature of the subject. The field is filled in only for verbs in which the semantics of the subject (described by one of the six possible values mentioned above) helps disambiguate the proper English equivalent. Some examples are given in *Table 14*.

Polish Verb	Semantics of subject	English Equivalent
chodzić	Animate	walk
	Inanimate	work
zwracać się	Animate	turn
	Inanimate	pay off
pochodzić	Human	come from
	Animate (non-human)	be descended
	Inanimate	originate, derive

Table 14

6. Application of the *Context* field

Context is handled in the system in a non-sophisticated manner. A list of about 20 contexts (related to computer science) has been created. For each context some characteristic words have been chosen from the dictionary which set the context in a text. For example, *drukarka (a printer)* sets the context for *printing*, *joystick* sets the context for *games*. Such characteristic words have been described in the *Context* field by *+context* (e.g. *+games* for joystick). Entries which may have different English equivalents according to context have been described by *?context*. Here are some examples:

Polish lexeme	Context	English Equivalent
wspólny	?hardware	shared
	no context	common
dołączać	?email	attach
	no context	join, add
ruch	?games	move
	?transport	traffic
	no context	movement, motion

Table 15

The algorithm treats context in a simple manner (if a word must be disambiguated by context, the algorithm looks for words assigning the context (marked in the dictionary by *+context*) in the neighbourhood. If no such words have been found, the context is set for *no context* (this approach requires any entry to have at least one default equivalent with no constraints set on the context).

7. Polish syntax

This field is consulted in the syntactical analysis. For example, for verbs, reflexiveness of the Polish part may be coded here. Some Polish verbs require a shift between subject and object when transferred into English. This type of information is also coded in the field. In order to translate a Polish sentence *Wystarcza mi danych* into *I have enough data* the algorithm needs the verb *wystarcza* to be denoted in a special way in the dictionary. Such information is coded in the field by *D → subj* (a Polish object in dative should be shifted into English subject). The modal verb *brak* has two different equivalents, each of them possessing a shift: D → subj for the equivalent *to lack* (e.g. *Brak nam funduszów → We lack funds*) and *G → subj* for the equivalent *be missing* (e.g. *Brak kilku osób → Several persons are missing).*

In paragraph 10 the field *EnglishSyntax*, which stores information on syntactical features of English equivalents, is discussed. The shift of syntax (which "lies in between" Polish syntax and English syntax) could as well be coded there. However, it is more convenient for the algorithm of the system POLENG to "know about the shift sooner".

The field *Polish Syntax* is relevant also for some adjectives and adjective pronouns. The predicative/attributive use may be coded there.

Polish Adj	Polish Syntax	English Adj	Description (Examples)
dodatkowy	attributive use (only)	extra	*dodatkowy procesor → extra processor*
	attributive or predicative use	additional, supplementary	*to złącze jest dodatkowe → this interface is supplementary*
sam	attributive use	very	*sam pomysł → the very idea*
	predicative use	alone, oneself	*przyszedł sam → he came alone; zrobił to sam → he did it himself*
niezły	predicative use	not bad	*on jest niezły → he is not bad*
	predicative or attributive use	pretty good	*To jest niezły komputer → This is a pretty good computer*

Table 16

8. Priority

One Polish lexeme may have a few English equivalents which are not distinguishable by the apparatus described above. The field *Priority* tells the algorithm the order in which the equivalents should be substituted. The order is assigned by the lexicographer on the basis of lexicographical sources, available text corpus (the frequency of occurrence being the criterion for establishing the order) or his/her intuition.

The equivalents in the field may be (almost) synonymous, e. g. *proper, suitable, appropriate* (all being the translations of the Polish adjective *właściwy*) or have different meanings which have not been disambiguated by other fields (e.g. *comment, remark, attention* -- equivalents of the Polish noun *uwaga*).

The output of the translation system includes equivalents of the highest priority. The user may choose between other equivalents in a convenient way.

9. English Lexeme Identifier, English Inflection Code

The fields contain the canonical form of an English equivalent and the code of its inflection respectively. The fields are necessary for the algorithm to generate appropriate inflected forms of the English output.

10. English syntax

The field is consulted by the translation algorithm in the process of syntactical synthethis. For verbs, the field describes features (not mutually exclusive) exemplified in *Table 17*.

Polish Verb	English Verb	Feature	Description (Examples)
sądzić	think	stative	*Teraz sądzę, że masz rację → Now I think you are right*
widzieć	see	"can"	*Widzę cię → I can see you*
dostosować się	adapt oneself	reflexive	
włączyć	switch on	"T2"	*Włącz to → Switch it on*

Table 17

The feature described in the above table as "can" concerns verbs of perception which tend to be preceded by the auxiliary verb *can*. The feature described as "T2" concerns phrasal verbs in which the complement may be placed before or after the preposition depending on its structure.

The field is relevant also for some adverbs and adverbial pronouns like *dużo, ile*. The value of the field (*plural* or *singular*) tells the algorithm to choose between the equivalents like *much, many* or *how much, how many* depending on the number of a nominal group following the adverbs.

11. Conclusion

The strategy of developing the *POLENG* system may be described as “step by step”. The research is centred on the applicability. All features described in the dictionary either have already been implemented in the system or the algorithm of implementing them has already been worked out.

The main drawback of such a pragmatic approach is that the system is imperfect by assumption. We believe that the format of the dictionary as well as the technology of its storage (described in [5]) is universal enough to allow for further improvement of the algorithm without the necessity of changing the formalism. One of several possible further directions of such improvement is presented in [3].

REFERENCES

[1] Jassem K., (1998), *Format wejścia w polsko-angielskim słowniku elektronicznym konstruowanym dla potrzeb tłumaczenia automatycznego*, w: W. Jassem, C. Basztura, K. Jassem: "Speech and Language Technology, vol. 2", Poznań

[2] Saloni Z., Świdziński M., (1998), *Składnia Współczesnego języka polskiego,* PWN, Warszawa

[3] Krynicki G., (1999), *Suggested Improvements in the Linguistic Aspect of the Electronic Polish-English Dictionary*, w: W. Jassem, C. Basztura, G. Demenko K. Jassem: "Speech and Language Technology, vol. 3", Poznań

[4] Jassem K., (1997), *POLENG - a Machine Translation System Based on an Electronic Dictionary*, w: W. Jassem, C. Basztura (eds) "Speech and Language Technology, vol. 1", Poznań 1997.

[5] Lison M., (1998), *Model i realizacja struktury danych leksykalnych słownika komputerowego skonstruowanego dla potrzeb automatycznego tłumaczenia*, w: W. Jassem, C. Basztura, K. Jassem: "Speech and Language Technology, vol. 2", Poznań